Preparation for Term Paper

(Summary Slides for Bloom et al. 2019)

Study Design

- **16 Parent Strains**
  - >1,000x deep sequencing
- **~13,950 recombinant haploid yeast segregants**, generated from crossing each parental strain to two other strains
  - Average of 872 progeny per cross
  - Measure growth in 38 environments (with duplicates)
  - Median of 2.3x coverage
- **32 "samples"** (one for each set of segregants produced from parental crosses) are deposited in PRJNA549760.
- **Code** and **processed data** has also been made publicly available.

Study Results

- 27.8% of biallelic variants are rare, but the median contribution of rare variants was 51.7%.
- That said, ***Figure 3*** shows some traits with a high fraction of variability explained by rare variants (in blue) but also some traits (Copper Sulfate, Lithium Chloride, etc.) that are ~100% explained by common variants (in gray).
  - Because the segregants genotypes are imputed from the much higher coverage parental genome sequences (and, by design, accurate imputations should be for one of the two parental strains used to generate those segregants), I assume that all variation considered is inherited (and there are no *de novo* variants being called or considered for impacting a trait).

Previous Evidence for Importance of Rare Variants

- Ehrenreich et al. 2012 – this study indicates "*variants that contribute to trait variation are shifted to lower frequencies when compared to all sequence variants*."

- General review in Gibson 2012.

- In the discussion, Bloom et al. 2013 mentions that rare variants may play a greater role in human genetics while also mentioning "*[because] all alleles are fixed at a frequency of one-half in a cross, we cannot yet delineate the contributions of common and rare variants to inherited variation, but we plan to do so in future studies*". So, I believe that is part of the rationale for this study.

Genetic Interactions Versus Additive Effects? (**Variance Components Model**)

- Text references Bloom et al. 2015, de los Campos et al. 2015, and Yang et al 2010.
- This may relate to code within *variance_components_by_AF.R* and/or *variance_components_within_cross.R*?

Fine-Mapping Identification of Causal Genes

- Posterior Probability of Causality (PPC)
  - I believe this relates to code within *QTL_causality.R*, where I can identify the previously characterized results for validation:

```
# Figure 4 ------------------------------------------------------------------
CG1=ggplot(tdf, aes(x=Rank, y= pCausalSum))+ theme_classic()+ylab('PPC')+
    geom_point(size=2) +
    geom_label_repel(force=1.5,
                     direction='both',
                     aes(label=NAME),
                     segment.color = 'black',
                     segment.alpha =.25 )+
    geom_point(data=tdf[tdf$previously_identified,],color='red')+
    theme(axis.text.x=element_text(size=rel(1),color='black'),
          axis.text.y=element_text(size=rel(1),color='black'))
```

- Farh et al. 2015 describe a Bayesian approach to identify causal SNPs.
  - Greater detail is provided in the Methods within the "*Probabilistic identification of causal SNPs (PICS)*" section.
  - Assuming the causal variant is a SNP, this strategy relates to permutations that I believe assist in the calculation of Bayesian probabilities for various SNPs.
  - For example, this would be different than determining a causal SNP based upon a loss-of-function prediction and expectation/prediction of gene function.
- ***Figure 4*** (in this study) includes experimentally validated variants among with all variants identified with FDR < 0.20.

Previously Characterized Loci (from *QTL_causality.R*)

```
prev.mapped=read.xls('/data/rrv2/genotyping/RData/NIHMS544073-supplement-01.xls', pattern='Table S1')
pm=unique(as.character(prev.mapped[-1,1]))[-c(96,97)][-16]
# HO is not segregating here and signal we see is likely due to effects of resistance casettes integrated at HO
pm=pm[pm!='HO']
#R> unique(QTGsorted.resolved$NAME)[which(unique(QTGsorted.resolved$NAME) %in% pm)]
# [1] "PCA1"  "RPI1"  "CYS4"  "HO"    "PHO84" "PDR5"  "GAL3"  "CAT5"  "IRA2"
#[10] "IRA1"  "MKT1"  "END3"  "FLO11" "SAL1"  "CYR1"  "TAO3"  "RGA1"  "HAP1"
#18
# ??? still missing
# SWH1 (Wang et al.)
# TOR1 and WHI2 (Treusch et al.)
# ENA1 (Steinmetz) ???
# ENA5 (?)
# KRE33 (desai)
# PMR1 (us)
# MAl11 Lit         ???
pm2=c(pm, 'SWH1', 'TOR1', 'WHI2', 'ENA1', 'ENA5', 'KRE33', 'PMR1', 'MAL11')
unique(QTGsorted.resolved$NAME)[which(unique(QTGsorted.resolved$NAME) %in% pm2)]

# enrichment of known genes at top of list
qtgrs=QTGsorted.resolved
qtgrs$previously_identified=qtgrs$NAME %in% pm2
qtgdf=data.frame(qtgrs)
```

Joint QTL Mapping

```
### ---Joint QTL mapping analysis ------------------------------------------------------------------------------------------------
#
#pre-processing
parents.list=lapply(parents.list, function(x) {
                  z=x;
                  z$marker.name.n=paste0(z$marker.name, '_', seq(1:nrow(z)))
                  return(z) })

# load JS 1,011 isolates allele frequency data into a structure
js.rr.overlap.allele.frequencies='/data/rrv2/1002genomes/isec_ouput/out.frq.mod'
sacCer3_CBS432_alignment.variants='/data/rrv2/spar_alignment/filt.snps'
sacCer3_CBS432_alignment.coords='/data/rrv2/spar_alignment/out.mcoords'
iseq.freqs=buildJS_variants_annotation_table(js.rr.overlap.allele.frequencies,sacCer3_CBS432_alignment.variants,sacCer3_CBS432_alignment.coords)
#save(iseq.freqs, file='/data/rrv2/genotyping/RData/iseq.freqs.RData')
load('/data/rrv2/genotyping/RData/iseq.freqs.RData')

# do multi-cross analysis using all called variants
#jointPeaks5=mapJointQTLs(n.perm=1000, FDR.thresh=.05, parents.list, pheno.resids, seg.recoded)
#load('/data/rrv2/genotyping/RData/jointPeaks5.RData')
#jP=rbindlist(jointPeaks5, idcol='chromosome')
#jPs=split(jP, jP$trait)
#save(jointPeaks5, file='/data/rrv2/genotyping/RData/jointPeaks5.RData')

# do multi-cross analysis using 1,011 panel variant data only
jointPeaksJS=mapJointQTLsJS_variants(n.perm=1e3, FDR_thresh=.05, parents.list, pheno.resids, seg.recoded, iseq.freqs, filterJS=T)
#save(jointPeaksJS, file='/data/rrv2/genotyping/RData/jointPeaksJS.RData')
```

Overall Goals

1) Identify analysis that can be completed within ~1 month.
   - I believe the underlying data used to produce ***Figure 3a*** should be helpful.
     - In particular, there are summary statistics in *elife-49212-fig2-data1-v2.xlsx*.
   - It is possible that this may be reduced to *one trait* and *one set of segregants* from a biparental cross.
2) Perform analysis similar to the *R/qtl2* user guide.
3) If possible, check ranking of validated results with different methods.
   - I believe the *red points* defined in ***Figure 4b*** should be helpful.
     - In particular, there are summary statistics in *elife-49212-fig4-data1-v2.xlsx*.

Additional QTL Notes

Jia Lab Publication Notes for *BPSC 234*

- I looked at the publications from the Jia Lab website to help complement the course material and term paper preparation.
  - I considered for QTL methods to consider for this project from the following publications:
    - Herniter et al. 2019 – describes the use the following R packages:
      - R/*mpMap* (Huang and George 2011 – now available as *mpMap2)*
      - R/*qtl* (Broman et al. 2003 : there is also Arends et al. 2010 and **Broman et al. 2019** publications)
        - ***R/qtl2*** has a **separate repository**.
      - Text describes *snow* used as described Herniter et al. 2018, with *cal.genoprob()* function.
        - However, I think this may be ***calc.genoprob()*** from ***R/qtl*** (or *calc_geneprob()* from *R/qtl2*)?
    - Wang et al. 2024 – used *GEMMA* (Zhou and Stephens 2012) for analysis
    - Ashworth et al. 2019 – used Interval Mapping (*IM*) and Kruskal-Wallis (*KW*) tests within *MapQTL* (version 5 for publication, currently version 7) for analysis.
  - Additional multi-omic methods were discussed in Wang et al. 2019. This includes *BLUP* and *LASSO*, among others.
  - Additional multi-locus methods were discussed in Zhang et al. 2019 . This includes *mrMLM/FASTmRMLM* and *GAPIT*, among others.
  - However, I thought the other methods (***R/qtl2***, ***GEMMA***, ***MapQTL***, **custom code for *Bloom et al. 2019***, etc.) may be a better fit for using this data to start to learn more about high-throughput QTL analysis.
- I then looked for some publications that cite *R/qtl* and/or *GEMMA.*
  - This includes Uffelmann et al. 2021, Nadeem et al. 2017, Peterson et al. 2012, Pérez and de los Campos 2014, Lee et al. 2014, Huang et al. 2018, etc.
  - There are also scripts for this publication that use the R *qtl* library (such as *mapping.R, simulateArchitecture.R,* and *variance_components_within_cross.R*). The *rrBLUP* library was also used, and R/qlt2 was referenced in *R_dependencies.csv*.
  - If needed, publications citing methods within *MapQTL* can also be checked.